VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Reliability of lumbar spine radiograph reading by chiropractors

Assendelft, W.J.J.; Bouter, L.M.; Knipschild, P.G.; Wilmink, J.T.

published in
Spine
1997

DOI (link to publisher)
10.1097/00007632-199706010-00013

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Assendelft, W. J. J., Bouter, L. M., Knipschild, P. G., & Wilmink, J. T. (1997). Reliability of lumbar spine radiograph reading by chiropractors. *Spine*, *22*(11), 1235-1241. https://doi.org/10.1097/00007632-199706010-00013

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 05. May. 2023

SPINE Volume 22, Number 11, pp 1235–1241
©1997, Lippincott–Raven Publishers

Reliability of Lumbar Spine Radiograph Reading by Chiropractors

Willem J. J. Assendelft, MD, PhD,* Lex M. Bouter, PhD,*†
Paul G. Knipschild, MD, PhD,‡ and Jan T. Wilmink, MD, PhD§

Study Design. An intraobserver and interobserver study on the reproducibility of data was performed.

Objective. This study investigates the variability in the interpretation of lumbar spine radiographs by chiropractors working in private practice.

Summary of Background Data. In chiropractic practice radiographs are used often, but this use is currently under debate. Therefore, there is a need for further study of the value of diagnoses made by radiographs by chiropractors. An acceptable intra- and interobserver agreement in radiograph reading is a prerequisite for a useful application of radiographs as a diagnostic tool in daily practice and in research.

Methods. Four chiropractors read 100 blinded sets of standard, erect anteroposterior and lateral lumbar radiographs independently. The same set was read in two separate sessions with a 2-month interval. The first session revealed the interobserver agreement. The comparison of the ratings by the same assessor in the two sessions indicated the intraobserver agreement. The assessors used a specially developed criteria list with emphasis on "nonspecific" radiographic findings. The prevalence of some important categories was increased artificially. Agreement was expressed in percentage agreement and generalized kappa, combining the results of all four assessors.

Results. Most kappas ranged from 0.40 to 0.75, representing fair to good agreement. In general, intraobserver agreement was better than interobserver agreement. The low kappas that were found may be explained partially by the high-agreement–low-kappa paradox as a result of a low prevalence.

Conclusion. The kappas and percentage agreement were acceptable, although not excellent. These results will be beneficial for future research on the value of radiograph diagnosis of nonspecific findings for delivery of safe and effective chiropractic therapy. [Key words: chiropractic, lumbar vertebrae (radiography), observer variation, reproducibility of results, spinal diseases (radiography)] **Spine 1997;22:1235–1241**

From the *Institute for Research in Extramural Medicine, Vrije Universiteit, Amsterdam, †Department of Epidemiology and Biostatistics, Vrije Universiteit, Amsterdam, ‡Department of Epidemiology, Maastricht University, Maastricht, and §Department of Radiology, Academic Hospital Maastricht.
Acknowledgment date: February 1, 1996.
First revision date: September 9, 1996.
Second revision date: January 8, 1997.
Acceptance date: January 11, 1997.
Device status category: 1.

Just 15 years after the discovery of x-rays, radiography already was included by chiropractors in their methods of examination.[25] At present, plain radiograph readings are still an important component of the chiropractic diagnostic working method. For example, nearly 60% of the practices in the Netherlands and 80% of the practices in the United States have their own radiograph capabilities.[2,4] In a recent survey among chiropractors in the Netherlands, almost all chiropractors stated that direct access to x-ray equipment was essential.[4] The importance of plain film radiography was a subject of the 1993 Mercy Condor Consensus Conference, an expert panel of the leading educators and researchers in the chiropractic field. Plain film radiography was rated as an established element of practice, and the scientific evidence of diagnostic value was rated to be high.[3]

In the medical field, the use of plain radiographs of the lumbar spine at present is restricted mainly to the exclusion of severe pathology, such as malignancy, inflammatory spondylarthropathy, or fracture.[14] These diagnoses generally are labelled as "specific causes" of low back pain. The clinical relevance of the identification of putative "nonspecific causes" of low back pain (the more prevalent congenital, structural, and degenerative abnormalities) is considered to be negligible.[43] Therefore, on the basis of previously published checklists, recent guidelines restrict the use of plain radiographs of the low back to a small, well-defined category of patients who have an increased risk for an underlying specific cause.[5,14,28,36,37] This restricted indication also is advocated by some leading chiropractic authors, whereas others in the chiropractic profession strongly oppose this point of view and advocate a much wider use.[32,33,39,45] According to these chiropractors, entities such as scoliosis, traction osteophytes, and disc degeneration, which in the medical literature are regarded as "nonspecific" findings, may alter the locus or direction of the manipulative intervention and identify contraindications for specific techniques. Whether this leads to increased safety and effectiveness in chiropractic has not yet been studied.[23,24]

The first step in the evaluation of diagnostic testing is knowledge about the reliability of the results.[38] The most cited chiropractic reliability study, by Phillips et al,[31] revealed that relatively few items had a high inter-rater reliability.[20] Another authoritative publication by Deyo

et al[13] concluded that among trained medical radiologists "agreement was less than optimal for most therapeutical diagnoses."

It is questionable whether the skepticism about the reliability of plain radiograph reading is yet sufficiently justified. Firstly, the articles by Phillips et al[31] and Deyo et al[13] reveal that the agreement was acceptable on several items that are of special relevance to chiropractic practice (*i.e.*, spondylolisthesis, short-leg discrepancy). Secondly, acceptable levels of agreement in recent medical studies have been reported.[35,41] Thirdly, the research methodology of a reliability study greatly influences its final results. The appropriate choice of statistics is of the utmost importance. Haas[21] advises the use of kappa for nominal data. Involvement of more than two assessors makes the procedure less sensitive for assessors with a controversial method of scoring. A generalized kappa, combining the results of all assessors, will smooth the impact of a controversial assessor.[21] Most available reliability studies use an extensive criteria list with many items on a random sample of radiographs (for example, Phillips et al[31] scored on 53 items), including entities with a relatively low prevalence. In these cases there is a great possibility that the error of "high agreement–low kappa" occurs because agreement is limited mainly to only one of the possible rating choices.[15,21] In addition, a small number of observations for a certain phenomenon renders a less stable estimate of observer agreement.[13] In the design of a reliability study this can be prevented by a deliberate over-representation of phenomena that are claimed by chiropractors to be of importance. A recent study by Davies et al[11] is an example of a study using this procedure to optimize efficiency.

The present study aimed to assess the inter- and intraobserver reliability of the reading of lumbar spine radiographs in chiropractic practice and tried to avoid the shortcomings mentioned above. To ensure an optimal estimate of agreement, the results are expressed in both a generalized kappa and the percentage of agreement. In the selection of the radiograph set, a sufficiently high prevalence of nonspecific entities considered to be relevant for chiropractic practice was established.

■ Methods

Rating List. A committee of the Netherlands Chiropractors' Association established a rating list for lumbar radiographs based on a list for computerized practices. For criteria that applied to several anatomic levels, only the lower three levels were included in the list to limit the amount of time spent per radiograph (resulting in ratings for L3–L4–L5, or L3–L4, L4—L5, and L5–S1). A detailed operationalization was written for every criterion. In the choice of the items, special emphasis was laid on the relevance to chiropractic practice. More details are provided in the Appendix.

Radiographs. One hundred erect anteroposterior and lateral radiographs of the lumbar spine were selected mainly from a 1-year sample of the files of a radiology service for general practitioners. The age of the patients involved was 18 years and older, and radiographs with osteosynthesis, prostheses, laminectomies, and other remarkable features not of interest to the study were excluded. In the selection of the radiograph set the authors aimed at an over-representation of scoliosis, degeneration, spondylolisthesis, spina bifida, transitional anomalies, osteoporosis, tumors, and pathologic abdominal arteries, to ensure a sufficient number of abnormalities in the various categories. Five radiographs with tumors were added from the hospital files of the radiology department. The radiographs were copied and blinded for all data, such as name and date of birth. Pelvic, lumbosacral spot, and oblique radiographs, when present, were not included in the set. A local chiropractor, who was not involved in the remainder of the study, checked the technical quality of the copies. Radiographs of insufficient quality were copied again, and if the quality remained insufficient, the radiograph was replaced by another one. To enable accurate scoring of the anatomic levels of abnormalities, the first lumbar vertebra was marked on every radiograph in the series.

Assessment. The assessment was made by four chiropractors (subsequently referred to as A, B, C, and D), who all had more than 3 years of experience in private practice and had graduated from the same college. Before the actual reading commenced, two consensus meetings were held to ensure uniform interpretation of the rating list. The radiographs were divided into subsets of 12 to 16 radiographs, and in the viewing room of a radiology department one subset per session was read by all chiropractors at the same time. After a 2-month interval the same set was read again in a random order, thus for each assessor made two ratings for each radiograph (rounds I and II).

Analysis. The prevalence of an abnormality was calculated from the average rating of all four assessors on both ratings. As indicators for the level of agreement Cohen's kappa and percentage agreement were calculated. Kappa is an index that provides an indication of agreement beyond chance.[7,8,38] A value of 0 indicates no agreement beyond chance, and 1 means perfect agreement.[7,38] The comparison among the ratings of the four assessors in round I provided the interobserver agreement. The comparison between the ratings by the same assessor in rounds I and II provided the intraobserver agreement. The kappa and percentage agreements first were calculated for all possible pairs in SPSS-PC 5.0 (SPSS, Inc., Chicago, IL) generating data for six pairs for the interobserver agreement (pairs A–B, A–C, A–D, B–C, B–D, C–D) and four pairs for the intraobserver agreement (A–A, B–B, C–C, D–D). The generalized kappa and percentage agreement summarizing the results of the six pairs for the interobserver agreement and the four pairs for the intraobserver agreement, respectively, were calculated using the formulas for the calculation of kappas and corresponding standard errors for more than two raters of Fleiss et al.[16,18,19] The generalized kappa, with corresponding 95% confidence interval, and percentage agreement for items that applied to several anatomic levels were calculated for the subitems for three separate levels and for all three levels combined.[1]

■ Results

The generalized kappas and percentage agreement are presented in Table 1. (The data for all pairs of assessors

Table 1. Inter- and Intraobserver Agreement in the Interpretation of Lumbar Spine Radiographs (nonrandom, selected set, N = 100)*

Item§	No. of Categories	Prevalence‖	Interobserver Agreement†		Intraobserver Agreement‡	
			Percentage Agreement	Kappa (95% CI)	Percentage Agreement	Kappa (95% CI)
Nonspecific: 3 levels combined						
Corpus deviation	3	54%	76%	0.58 (0.55–0.61)	85%	0.74 (0.70–0.78)
Body inferior inclination	3	53%	82%	0.71 (0.68–0.75)	87%	0.79 (0.75–0.84)
Space narrowing	2	46%	79%	0.58 (0.53–0.63)	85%	0.69 (0.63–0.75)
Spondylosis	2	28%	83%	0.57 (0.52–0.62)	83%	0.71 (0.65–0.76)
Facet artrosis	2	25%	69%	0.17 (0.13–0.22)	82%	0.53 (0.47–0.59)
Spondylolisthesis	3	8%	92%	0.66 (0.62–0.70)	94%	0.71 (0.67–0.75)
Vacuum sign	2	2%	98%	0.55 (0.45–0.61)	99%	0.75 (0.67–0.83)
Laterolisthesis	2	1%	98%	0.32 (0.20–0.38)	99%	0.62 (0.53–0.71)
Nonspecific: single items						
Gravity line L3	3	56%	77%	0.61 (0.55–0.67)	84%	0.74 (0.66–0.82)
Scoliosis	3	51%	65%	0.46 (0.40–0.52)	85%	0.76 (0.69–0.83)
Spinous torsion	3	49%	73%	0.56 (0.52–0.62)	83%	0.73 (0.66–0.80)
Lumbar lordosis	3	42%	53%	0.20 (0.14–0.26)	77%	0.59 (0.51–0.66)
S base inferior inclination	3	22%	80%	0.40 (0.34–0.46)	84%	0.55 (0.47–0.62)
S base angle	3	20%	79%	0.41 (0.35–0.48)	89%	0.67 (0.59–0.75)
SI degeneration	2	18%	70%	0.23 (0.17–0.29)	79%	0.35 (0.28–0.42)
Transitional anomaly	2	11%	87%	0.41 (0.35–0.46)	94%	0.71 (0.64–0.78)
Spina bifida	2	11%	93%	0.68 (0.60–0.76)	94%	0.70 (0.60–0.80)
S spinous rotation	3	7%	87%	0.26 (0.18–0.33)	92%	0.40 (0.30–0.49)
SI inflammatory	2	2%	88%	0.17 (0.05–0.30)	91%	0.28 (0.11–0.44)
Schmorl's nodes	2	0%	100%	0.00 -	100%	0.00 -
Specific: single items						
Osteoporosis	2	24%	83%	0.55 (0.47–0.63)	87%	0.65 (0.55–0.75)
Pathological abdominal arteries	2	25%	94%	0.84 (0.76–0.92)	94%	0.83 (0.73–0.93)
Fractures (old and fresh)	2	8%	90%	0.51 (0.41–0.61)	97%	0.72 (0.60–0.84)
Tumor (benign and malignant)	2	5%	95%	0.46 (0.37–0.54)	95%	0.42 (0.32–0.52)
Infection	2	0%	100%	0.00 -	100%	0.00 -
Syndesmophytes	2	0%	95%	0.26 (0.18–0.30)	97%	0.55 (0.45–0.65)
Paget's	2	0%	100%	0.00 -	100%	0.00 -

* Ranked in order of prevalence.
† Combines the results of six pairs of assessors.
‡ Combines the scores for four assessors.
§ Lower anatomic levels, depending on item: L3, L4, L5, or Le–L4, L4–L5, L5–S1.
‖ Prevalence calculated as the average prevalence of abnormalities by all 4 assessors in the 2 assessment rounds.
95% CI = 95% confidence interval; S = sacral; L = lumbar; SI = sacroiliac.

and for the subitems covering the separate anatomic levels, as initially calculated, are available on request from the first author [WJJA]). As expected the intraobserver agreement was consistently higher than the interobserver agreement.

Although the prevalence of abnormalities relevant to chiropractors in the radiograph set was manipulated, the authors did not succeed in establishing sufficiently high prevalence for several items. There was an inverse relation between prevalence and the percentage agreement: on items with a low prevalence (≤ 10%) the percentage agreement was generally high; for items with a high agreement, however, the kappa can be paradoxically low.[15] Therefore, for items with a high agreement but a low kappa, more emphasis should be laid on the percentage agreement. The items in the upper portion of the table ("nonspecific; three levels combined") were assessed initially as subitems for several anatomic levels separately and later synthesized to a generalized kappa and percentage agreement for all levels combined. The prevalences reported for these items are the average of all three levels. The items presented in the lower part of the table ("nonspecific; single item" and "specific; single item") cover items that were assessed for the entire lumbar region.

For the various subitems, the kappas and percentage agreement did not vary substantially. Therefore only the generalized kappas and percentage agreement are presented. The only exception was spondylolisthesis, which clearly had a better agreement at L5–S1. The interobserver percentage agreement and kappa for L3–L4, L4–L5 and L5–S1 were 92% and 0.37, 94% and 0.62, and 90% and 0.74, respectively, and the intraobserver agreement was 95% and 0.50, 94% and 0.68, and 92% and 0.77, respectively. For this item the prevalence at L5–S1 differed substantially from the prevalence at the two other levels (L3–L4 and L4–L5), namely 18% compared with 6% and 1%, respectively (note: the 8% prevalence in Table 1 is the numerical average of the preva-

lences on these separate anatomic levels; the prevalence of spondylolisthesis on either level, however, is the sum of these percentages, namely 25%).

■ Discussion

This study aimed to evaluate the reliability of radiograph reading in chiropractic practice. Therefore, chiropractors without any additional qualification in radiology and who were working full-time in private practice were selected. Two half-day sessions were used to enable them to become familiar with the operationalization of the rating list. However, no extensive consensus meetings were held. The authors of this study aimed to include a sufficiently high prevalence of most abnormalities relevant to chiropractors in the rating list and were able to achieve this for most items. In the authors' opinion, the simultaneous presentation of the prevalence, kappa, and percentage agreement enables adequate judgment of the results. Combining the results of all four assessors provides a balanced reflection of the agreement.

Kappa is dependent on the prevalence, the number of categories, possible weighing in the case of more than two categories, and the presence of bias.[7] It is not simple, therefore, to assign a definite interpretation to kappa values.[7] A frequently cited standard interpretation is that of Landis and Koch[27], which was adapted by Fleiss. According to Fleiss,[19] values of ≤ 0.40 represent poor agreement, values between 0.40 and 0.75 represent fair to good agreement, and values > 0.75 indicate excellent agreement.[19] Most kappas in the present study can be classified, therefore, as indicating fair to good agreement. As could be expected, for the most part better intraobserver agreement than interobserver agreement was found.

The low kappas that were found may be explained partially by the high-agreement–low-kappa paradox as a result of low prevalence.[15] One clear exception is facet arthrosis, which produced a remarkably low interobserver kappa (0.17), despite an adequately high prevalence (25%). This is also a consistent finding in other studies. For plain radiography, low kappas for lumbar facet joint abnormalities also were found by Pathria[30] (kappa 0.26 for facet joint arthritis), Coste et al[9] (kappa for facet abnormalities, 0.16–0.31) and Deyo et al[13] (kappa 0.33 for joint sclerosis and 0.24 for joint narrowing). Also, computed tomography by Coste et al[10] showed low kappas for lumbar facet joint arthritis. In the present study, however, the difference with the corresponding intraobserver kappa (0.53) also indicates that for this item a more extensive consensus procedure might be helpful. The results from the study by Dehnugara[12] also point in this direction: the two chiropractic radiology experts who already worked closely together reached a kappa of 0.64 on facet joint degeneration. The low kappas on sacroiliac (SI) degeneration and SI inflammation are at least partly explained by the fact that on most of the radiographs the SI joints were not completely and, according to the chiropractors, not sufficiently well visualized.

Some aspects in the design and execution of the present study may have influenced the results in an adverse direction. Copies of the radiographs were used to blind them. In a number of radiographs there was a reduction in quality of the radiographs after the copying. Eventually, 12 of the 100 radiographs were considered to be of low quality by at least two of the four assessors. This poor quality most likely caused lower estimates of agreement. Despite sincere attempts, because of low numbers and quality problems the authors were not able to select enough cases of specific causes of back pain (tumors, infections, Paget's, and fractures) to produce a stable estimate of kappa. Arbitrarily, a prevalence of 10% can be considered to be a minimum requirement.[42] As a result, the kappa values for these specific diagnoses should be interpreted with great caution. Disagreement is more common when there are many diagnostic categories to consider and when abnormalities are mild.[13] In the interpretation of lumbar spine radiographs with an extensive criteria list, as in the present study, both conditions usually apply, which increases the likelihood of disagreement. The positioning of the patients and the way anatomic structures are projected may differ between the medical radiographs used in the set in this study and radiographs made in chiropractic practices. However, most Dutch chiropractors accept recently made medical radiographs of their patients as a part of their initial patient assessment.[4] Therefore, they should be accustomed to reading radiographs from various hospitals. However, better agreement might be attainable with radiographs from chiropractic practices. The authors of this study only had two half-day consensus meetings, mainly focussing on the interpretation of the criteria list. A medical study on lumbar radiograph reading showed that a joint meeting of the independent assessors after a series of radiograph readings, focussing on sources of disagreement, could substantially increase the level of agreement (kappa on a four-point scale increased from 0.49 to 0.69).[41] Perhaps a more extensive preceding consensus and feedback procedure would have increased the agreement, especially because the participating chiropractors are not working together in daily practice.

When compared with daily practice, some other study-related factors might have influenced the results of this study in a positive way. The authors decided to mark the first lumbar vertebra because in the case of a transitional vertebra there might be confusion about how to indicate the anatomic level. If the assessor had been mistaken about the proper anatomic level of a transitional vertebra, all items that had to be assessed for various anatomic levels ("nonspecific; single items") would have been assessed incorrectly, which the authors considered a penalty too large for this possible confusion. However, the authors believe that the impact of applying this mark

on the overall conclusions is relatively small. In addition, a standardized criteria list, the consensus meeting, the exclusion of a small number of technically inferior radiographs, and the fact that all four assessors were trained at the same college might have increased the level of agreement.

From the previous considerations it is clear that numerous factors influence the level of agreement, expressed as kappa or percentage agreement. Apart from the difficulty mentioned above in providing a definite comparison standard for kappa values, the comparison with other studies is also dubious. Dehnugara[12] applied the same study design (over-representation of specific findings) and a more or less similar checklist (32 items) as were used in the present study. The assessors were trained chiropractic radiologists. In general, her study revealed the same pattern of results as those in the present study: kappas indicating a fair to moderate interobserver agreement on most of the nonspecific findings that had a sufficiently high prevalence and low kappas and high percentages agreement for the less prevalent specific findings, such as lytic lesions (most likely tumors). In the study by Phillips et al,[31] a 56-item checklist (all nonspecific items) was used by three chiropractic assessors.[20] The results of the four-category items were expressed in intraclass correlation coefficients (ICC).[7] The authors considered 0.8 and above as good agreement and 0.6–0.8 as fair agreement. Prevalences were not presented in this publication. Six items scored a good agreement, and 16 scored fair agreement. The remaining 34 had an ICC lower than 0.6, and were considered to indicate poor agreement. The ICC can be regarded as a kappa with quadratic weights.[17] Consequently, it seems that the classification of the results in this study has been somewhat rigorous when compared with the classification of Fleiss.[19] In addition, agreement was taken to be acceptable on several items that are of special relevance to chiropractic practice (*i.e.*, spondylolisthesis and short-leg discrepancy). The other less recent chiropractic studies evaluating the reliability of the assessment of static lumbar radiographs are either flawed in design or have applied inappropriate statistics.[22]

The reliability of the interpretation of lumbar spine radiographs also has been studied for medical professionals. In the study by Deyo et al[13], 100 radiographs of an unselected population of patients with low back pain were assessed by two board-certified radiologists. Specific causes of back pain were excluded. They concluded that there were moderate to substantial levels of agreement for many findings including osteophytes, osteopenia, compression fractures, and spondylolysis. Agreement was only fair, however, for various changes in the facet joints (which were already discussed in this report), sacroiliac sclerosis, and transitional anomalies. These findings are considerably similarity to those of the present study. Deyo et al[13] concluded that their findings were "neither surprising, nor alarming." Coste et al[9] repeated the study of Deyo et al,[13] this time with rheumatologists as assessors. As in the present study, the highest levels of interobserver agreement were observed for disc abnormalities. There was only poor to slight agreement for facet joint abnormalities. There was more agreement on the presence of transitional vertebrae than was found in the present study (kappas 0.69 and 0.41, respectively). These authors concluded that "a significant variability in the interpretation was observed for many findings often considered important for benign low back pain." In the study by Riihimäki et al,[35] 50 radiographs from a population study were read by two radiologists. The weighted kappas for the four-category items were 0.42–0.88 for disc-space narrowing, 0.45–0.88 for anterior osteophytes (spondylosis), and 0.45–0.80 for end-plate sclerosis. They concluded that these findings, which are similar to those in the present study, were satisfactory. Evaluating the agreement on the Kellgren method for scoring disc degeneration Symmons et al[41] described a kappa of 0.49 for the first half of the reading and 0.69 for the second half, thus indicating the effect of training in the assessors. In the above-mentioned chiropractic and medical studies on the reliability of reading radiographs of the lumbar spine, the original authors of the various studies presented conflicting conclusions on more or less similar levels of agreement.

Kappas ranging from 0.40 to 0.75 (fair to good agreement) are not uncommon for advanced radiologic procedures such as computer tomography for lumbar disc hernia and lumbar facet osteoarthritis or magnetic resonance imaging of the lumbar spine or pelvis.[6,10,30,34,44] Outside the field of radiography, common diagnostic procedures such as palpation of the thyroid, palpation of the liver, or the detection of airflow obstruction produce kappas within the same range.[26,29,40]

In the present study most of the kappas and, in the case of low prevalence, the percentage agreement were acceptable, although not excellent. In future research the results of reading radiographs probably can be improved further. A consensus meeting with feedback on the scoring of a limited pilot set of radiographs also could further improve the results. Moreover, if possible, the set should consist of original radiographs, not copies. It is possible that the results can be improved even further if the positioning of the patient and the projection of anatomic structures were performed consistently according to chiropractic standards.

Another composition of the set of radiographs might provide information on aspects that could not be addressed in this study. A sufficient proportion of radiographs of specific findings (tumors, ankylosing spondylitis, infections, Paget's, and fractures) would provide more definite information on the reliability of chiropractic assessment of these features. On the other hand, the reading of a completely random set of radiographs (for example, primary care patients) would provide a true

estimate of the prevalence of nonspecific findings according to chiropractors.

This study solely evaluated the reliability of radiograph assessment. It remains unclear to what extent the identification of "nonspecific" findings alter the locus or direction of manipulation and how these "nonspecific" findings relate to contraindications for certain techniques. Given the frequent use of radiograph diagnosis, not only for exclusion of severe pathology, it is now up to the chiropractic profession to prove the additional value of radiograph diagnosis of nonspecific findings for the delivery of safe and effective therapy.[31,33]

Acknowledgements

The authors thank R. Blaauw, L. C. Kramer-Bakker and C. E. Pfeifle for their help in the preparation of the rating list, P. Keil for his help with the selection of the radiographs, and R. Bakker, J. A. Kuipers, R. Meijer, and M. Roosenberg (all members of the Netherlands Chiropractors' Association) for the many hours that they spent as assessors. The authors also thank M. E. Ooms for writing the computer program for the calculation of generalized kappas and percentages agreement.

References

1. Altman DG. Practical Statistics for Medical Research. London: Chapman & Hall, 1991:405–6.
2. Statistics on chiropractic offices and equipment. J Am Chiropr Assoc 1987;24:56–61.
3. Diagnostic imaging. In: Haldeman S, Chapman-Smith D, Petersen DM, eds. Guidelines for Chiropractic Quality Assurance and Practice Parameters. Proceedings of the Mercy Center Consensus Conference. Gaithersburg: Aspen, 1993:13–34.
4. Assendelft WJJ, Pfeifle CE, Bouter LM. Chiropractic in the Netherlands: A survey among Dutch chiropractors. J Manipulative Physiol Ther 1995;18:129–34.
5. Bigos S, ed. Acute Low Back Problems in Adults [report]. AHCPR Publication 95-0642. Rockville, MD: U.S. Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research, 1994.
6. Brant-Zawadzki MN, Jensen MC, Obuchowski N, Ross JS, Modic MT. Interobserver and intraobserver variability in interpretation of lumbar disc abnormalities. Spine 1995;20: 1257–64.
7. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. Br Med J 1992;304: 1491–4.
8. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
9. Coste J, Paolaggi JB, Spira A. Reliability of interpretation of lumbar spine radiographs in benign, mechanical low-back pain. Spine 1991;16:426–8.
10. Coste J, Judet O, Barre O, Siaud J, Cohen de Lara A, Paolaggi J. Inter- and intraobserver variability in the interpretation of computed tomography of the lumbar spine. J Clin Epidemiol 1994;47:375–81.
11. Davies AM, Fowler J, Tyrrell PNM, Millar JS, Leahy JF, Patel K, et al. Detection of significant abnormalities on lumbar spine radiographs. Br J Radiol 1993;66:37–43.
12. Dehnugara R. Inter- and Intra-rater Reliability in the Interpretation of Lumbar-Pelvic Radiographs [report]. Bournemouth: Anglo-European College of Chiropractic, 1994.
13. Deyo R, McNiesh LM, Cone RO. Observer variability in the interpretation of lumbar spine radiographs. Arthr Rheum 1985;28:1066–70.
14. Deyo RA, Diehl AK. Lumbar spine films in primary care: current use and effects of selective ordering. J Gen Intern Med 1986;1:20–5.
15. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol 1990; 43:543–9.
16. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378–82.
17. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educ Psychol Meas 1973;33:613–9.
18. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychol Bull 1979; 86:974–7.
19. Fleiss JL. Statistical Methods for Rates and Proportions. 2nd ed. New York: Wiley, 1981:212–36.
20. Frymoyer JW, Phillips RB, Newberg AH, McPherson BV. A comparative analysis of the interpretations of lumbar spine radiographs by chiropractors and medical doctors. Spine 1986; 11:1020–3.
21. Haas M. Statistical methodology for reliability studies. J Manipulative Physiol Ther 1991;14:119–32.
22. Haas M. The reliability of reliability. J Manipulative Physiol Ther 1991;14:199–208.
23. Haldeman S, Phillips RB. Spinal manipulative therapy in the management of low back pain. In: Frymoyer JW, ed. The Adult Spine: Principles and Practice. New York: Raven, 1991: 1581–1605.
24. Hall FM. Back pain and the radiologist. Radiology 1980; 137:861–3.
25. Hildebrandt RW. Chiropractic spinography and postural roentgenography: Part I: History of development. J Manipulative Physiol Ther 1980;3:87–92.
26. Holleman DR, Simel DL. Does the clinical examination predict airflow limitation? JAMA 1995;273:313–9.
27. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
28. Liang M, Komaroff L. Roentgenograms in primary care patients with acute low back pain: A cost-effectiveness analysis. Arch Intern Med 1982;142:1108–12.
29. Naylor CD. Physical examination of the liver. JAMA 1994;271:1859–65.
30. Pathria M, Sartoris DJ, Resnick D. Osteoarthritis of the facet joints: Accuracy of oblique radiographic assessment. Radiology 1987;164:227–30.
31. Phillips RB, Frymoyer JW, McPherson BV, Newburg AH. Low back pain: A radiographic enigma. J Manipulative Physiol Ther 1986;9:183–7.
32. Phillips RB. Plain film radiology in chiropractic. J Manipulative Physiol Ther 1992;15:47–50.
33. Plaugher G. The role of plain radiography in chiropractic clinical practice. Chiropr J Austr 1992;22:153–61.
34. Raininko R, Manninen H, Battié MC, Gibbons LE, Gill K, Fisher LD. Observer variability in the assessment of disc degen-

eration on magnetic resonance images of the lumbar and thoracic spine. Spine 1995;20:1029–35.
35. Riihimäki H, Wickström G, Hänninen K, Mattson T, Waris P, Zitting A. Radiographically detectable lumbar degenerative changes as risk indicators of back pain. Scand J Work Environ Health 1989;15:280–5.
36. Rosen M, ed. Back Pain. Report of a Clinical Standards Advisory Group Committee on Back Pain. London: Her Majesty's Stationary Office, 1994.
37. Royal College of Radiologists. Making the Best Use of a Department of Radiology: Guidelines for Doctors. London: Royal College of Radiologists, 1989.
38. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd ed. Boston: Little-Brown, 1991:24–31.
39. Schultz G, Phillips RB, Cooley J, et al. Diagnostic imaging of the spine in chiropractic practice: Recommendation for utilization. Chiropr J Austr 1992;22:141–52.
40. Siminoski K. Does this patient have a goiter? JAMA 1995; 273:813–7.
41. Symmons DP, Hemert AM van, Vandenbroucke JP, Valkenburg HA. A longitudinal study of back pain and radiological changes in the lumbar spines of middle aged women: II. Radiographic findings. Ann Rheum Dis 1991;50:162–6.
42. Triet EF van, Dekker J, Kerssens JJ, Curfs EChr. Reliability of the assessment of impairments and disabilities in survey research in the field of physical therapy. Int J Disabil Stud 1990; 12:61–5.
43. van Tulder MW, Assendelft WJJ, Koes BW, Bouter LM. Spinal X-ray findings and non-specific low back pain: A systematic review of observational studies. Spine 1997;22:427–34.
44. Wright JG, McCauley TR, Bell SM, McCarthy S. The reliability of radiologists' quality assessment of MR pelvic scans. J Comp Assist Tomograph 1992;16:592–6.
45. Wyatt L, Schultz G. The diagnostic efficacy of lumbar spine radiography: A review of the literature. In: Hodgsen M, ed. Current Topics in Chiropractic. Sunnyvale: Palmer College of Chiropractic-West, A2-2–A2-14.

Address reprint requests to

Willem J. J. Assendelft, MD, PhD
Institute for Research in Extramural Medicine,
Vrije Universiteit
Van der Boechorststraat 7
1081 BT Amsterdam
the Netherlands

■ Appendix. Operationalization of Some Criteria.

Spondylosis is based on the presence of corporal osteophytes.

Spondylolisthesis is only present if clearly visible or with a slipping of at least 2 mm.

Gravity line L3 is defined as the vertical line drawn down from the intersection of diagonal lines from the four corners of the corpus L3, not including possible osteophytes. "Normal" is intersection in the ventral one-third of the sacral surface, not including possible osteophytes. This line is drawn only if visually not obvious.

Scoliosis is judged only on its lumbar component. A curvature to one side with no clear return to midline can still be considered scoliosis.

Spinous torsion is considered to be rotation of two or more consecutive spinal segments.

Lumbar lordosis has three categories: curvature of less than 50°, between 50° and 60° and more than 60°. Lordosis is defined as the angle of the intersection of lines drawn through the inferior endplate of L1 and the superior surface of the sacrum, not including possible osteophytes. "Normal" is curvature between 50° and 60°.

Sacral base angle has three categories: less than 30°, between 30° and 50°, and more than 50°. "Normal" is 30°–50°.

Schmorl's nodes are short, abrupt impressions, usually covering only part of the distance of the endplate.

Pathologic abdominal arteries imply visible calcification of the vessel wall and/or aneurysm.

Fractures are not defined further as past or recent, consolidated or not. Obvious anterior wedging is considered to be a (compression) fracture.

Tumors are identified only as present or absent. Benign and malignant manifestations are both scored as tumors.

Infections of the vertebrae should have led to destruction of the surfaces of two congruent vertebrae and, most likely, joint narrowing. Sacroiliac infection will show sclerosis, (pseudo) widening of the joint surfaces, and/or irregular joint surfaces or fusion.